2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) | 979-8-3503-9874-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/SCEECS57921.2023.10063128

Deepfake Face Detection Using Deep InceptionNet Learning Algorithm

Prasannavenkatesan Theerthagiri
Department of Computer Science and Engineering,
GITAM University
Bengaluru, India,
*vprasann@gitam.edu,0000-0003-3420-598X

Ghouse basha Nagaladinne
Department of Computer Science and Engineering
GITAM University
Bengaluru, India
ghousebasha2029@gmail.com

***Abstract*—Deepfakes is digital manipulation techniques that use deep learning to produce deepfake (misleading) images and videos. Identifying deepfake images is the most difficult part of finding the original. Due to the increasing reputation of deep fakes, identifying original images and videos is more crucial to detect manipulated videos. This paper studies and experiments with different methods that can be used to detect fake and real images and videos. The Convolutional Neural Network (CNN) algorithm named InceptionNet has been used to identify deep fakes. A comparative analysis was performed in this work based on various convolutional Networks. This work uses the dataset from Kaggle with 401 videos of train sample and 3745 images were generated by augmentation process. The results were evaluated with the metrics like accuracy and confusion matrix. The results of the proposed model produces better results in terms of accuracy with 93% on identifying deep fake images and videos**

Keywords— Deepfake, Inception net, CNN(Convolutional Neural Network), Vision Transformers

I. INTRODUCTION

With the rise of smartphones and social media networks, deepfake videos have become very common. These gadgets have created fake news and videos, which are considered dangerous for society. Also, misleading images and videos are made by terrorist organizations to humiliate the people and world and threaten the nation [1]. An increase in virtualization and globalization made the world shrink but also invited some nonstate threats to the nation by using fake videos, radicalizing people from other religions, and propagating the agenda. Many high-profile people came under this trap and suffered from a lot of problems because of fake images and videos [2].

The face is the most distinctive feature of human beings. With the rapid advancement of face blends innovation, the security risk posed by face control is becoming increasingly critical. Human faces can frequently change by someone's look, which can show up as real and actual human faces because of many calculations that rely on profound acquiring innovation. It is a growing subset of counterfeit insights innovation in which anyone's face can match with someone's real face [3].

Deepfake substance is spreading faster than ever before in the twenty-first century. Because of the growing popularity of deepfakes, methods for detecting fake videos that are presented as real ones are becoming increasingly important. In this journal, we will look at other technologies that can be used to detect deepfake images [4].

In the last few decades, smartphone culture and the gradual growth of social networking sites have made images and videos digitally popular. This paper shows various types of vision transformers to perform inception net. It is used to determine the precision percentage and which method is more accurate and adequate for detecting deepfake or real videos [5].

II. LITERATURE SURVEY

An extensive literature study has been carried out on the related articles about deep fakes detection models and techniques for enhancing the existing approaches. A literature study is done on various data mining techniques

The following section describes the related papers which are studied in this paper.

Nishat Tasnim Roza et al. 2021 proposed a comparative analysis of the deepfake image detection method using a convolutional neural network. In their research, they described the following things. Human faces have very distinctive features in nature. Deepfake videos/images are revolutionary subduals of AI technology that use someone's face to overwrite someone's face. The deepfake images were not visible to normal eye vision due to the collapse of the pixel, skin tones, and facial shapes of images, which are artificial visual deformities. Not only images and videos but also audio can be converted into deepfake. Because of technological advancements, deepfake has become almost indistinguishable from natural images. As a result, people around the world are experiencing inevitable complications [1].

Kaipeng Zhang et al. 2016 proposed joint face detection and alignment using MTCN. In their research, they described the following things. Detection of fake faces and their alignment is critical to many face programs, consisting of face reputation and facial features analysis. However, the massive visible versions of faces, consisting of occlusions, massive pose versions, and excessive lighting, impose incredibly demanding situations for those obligations in actual global programs. The cascade face detector proposed Haar-Like functions and AdaBoost to teach cascaded

979-8-3503-9874-8/23/$31.00 ©2023 IEEE

classifiers, which obtain true overall performance with actual-time efficacy. However, some works suggest that this detector can also degrade considerably in actual-global programs with large visible versions of human faces despite extra superior functions and classifiers. Besides the cascade structure, introduce deformable element models (DPM) for face detection and reap splendid overall performance. However, they want excessive estimation rates and might require costly annotation within the schooling stage. Recently, convolutional neural networks have reaped splendid progress in many pc imaginative and prescient obligations, consisting of picture category and face reputation. Inspired by way of means of the best overall performance of CNN in pc imaginative and prescient obligations, several CNN-primarily based face detection strategies had been proposed in recent years. Yang et al. teach deep convolution neural networks for the characteristic facial reputation to acquire excessive reaction in face areas, yielding candidate home windows of faces. However, because of its complicated CNN structure, this technique is time pricey in practice. Li et al. use cascaded CNN for face detection; However, it calls for bounding field calibration from face detection with a greater estimation rate and ignores the inherent relation among facial landmarks localization and bounding regression [3].

Christian Szegedy et al. 2015 proposed Going Deeper with Convolutions. In their research, they described the following things. In the closing 3 years, they said their item class and detection competencies have dramatically progressed because of advances in deep getting to know about it. Statistics can be maximizing the development isn't always simply the end result of extra effective hardware, large data files, and larger fashions, However, particularly an effect of latest ideas, algorithms and progressed community architectures. No new statistics reasserts have been used, for example, via way of means of the pinnacle entries withinside the opposition except the identical class of opposition by using identification purposes. In their research, they said Google net has to be submitted to surely make use of 10 instances fewer parameters than the triumphing structure of Krizhevsky from years ago, even as being substantially extra accurate. On the item detection front, the largest profits have now no longer larger utility & larger network of deep, however from the deep architectures & old laptops, just same as the CNN set of rules via way of means of Girshick. Another first-rate issue along with the continuing of cellular and computed which is embedded, performance of the algorithm – in particular of the strength & reminiscence – profits. Concerns main are the layout deeply structure supplied on the note covered issue in place of maximum in the trial, the fashions have been made to maintain estimation finance by 1.6 multiplication of billions that provides by correct time, so that is used by now which is no longer grow to be a merely educational curiosity, however, will be placed to actual international use, even on huge data files, at an inexpensive price.

Here in the note, those can attention to green neural which is a deeper community structure in laptop vision, in community paper via way of means of Lin et al together with the well-known they want to move deeper net meme. In their case, the phrase deep is utilized special meanings: first of all, withinside the experience a brand stage of enterprise withinside additionally withinside the extra direct experience of multiplied community [2].

Thus, most of the literature survey focuses on the techniques to fetch data from Twitter and news articles after which this data is converted into the desired format and applied operations to get the intent of the user. However, they have not focused on the algorithm for classifying the topic of the news tweets.

III. METHODOLOGY

There are many methods to detect fake faces using GAN deepfake algorithm images, including conventional methods such as learning classifiers, DNN, CNN, and RNN. Besides these, there are some disadvantages of the existing system, i.e., most traditional convolutional neural network models lack precision. It uses face recognition methods which reduce the efficacy while extracting frames from the input video. The proposed system uses the inception net to analyze and identify the deepfake Images.

A. DEEPFAKE

It is a technique in points that must replace a targeted person in the confrontation at that time. It initially surfaced in the fall of 2017 as a script used to produce swapping faces adult content. A short while later, this strategy was improved upon by employing a small amount of code to effectively create a user-friendly piece of software known as a dummy app. The preparation of the autoencoders in parallel contains the fundamental idea. their design can change by the estimated production, necessary preparation time, anticipated quality, and available resources. A decoder network and an encoder arrangement are often chained together by an auto-encoder. The encoder's purpose is to reduce measurement error by encoding into a smaller number of elements of the statistics from the input layer. The decoder's objective at that moment is to implement these changeable to get a kind of estimate in initial gain.

Here A and B's aligned faces are collected to make deepfake photos. An auto-encoder EA is then trained to reproduce A's face from the data file of the human figure of A, and an automated encoder is trained by recreating B's face from the data file of human figure of B. the logic is to combine encoding mass of the two automated encoders EA and EB while isolating the corresponding decoders. After optimization, any image having an A-confrontation may be encoded using this common encoder and decoded by using the EB-decoder. Figures 1 and 2 explain this rule.

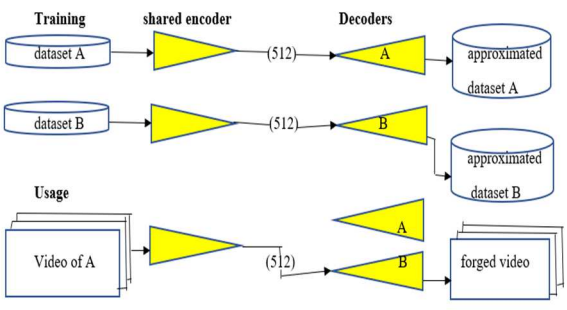

Fig: 1 Fake accounts principles

The prepared components with the identical encoder in yellow are the best. Foot: the part of the use process when pictures of A are translated using a decoder.

An idea for this strategy is to use an injector, which works by encoding common statistics such as light, location, and facial expression and recreating the usual distinctive forms of both individual figures. The important data can thus be separated from the morphological data on one side. Honing produces impressive results, which explains why the process is so popular. The final stage employs the provided clip, removing and changing the squared figure from each outline, and then generating using the modified auto-encoder.

DEEPFAKE ORIGINAL

Fig 2: The image on the (right) has been changed (left) by using the technique deepfake

This method works well for perfection. In essence, removing human facial contours may result in a large region with many versions of the same facial form. However, with increasingly sophisticated networks, such technological flaws may be easily avoided. More importantly, it is often true for other applications, such as autoencoders tend to reproduce fine points of interest poorly because they compress the input data into a small encoding area, resulting in a hazy result.

Although small details are better approximated, a larger encoding area is ineffective. Because, on the other hand, the approach to encounter loses authenticity because it frequently mimics the input confront, which may have an unfavorable effect because morphological statistics are provided to the decoder.

B. DATASET

An optional structure involves replacing the inception net's first two convolutional layers with a modified version of the Szegedy, inception module. The chapter's ideology is to expand workstations in which the demonstration is optimized by stacking the outgain of numerous component figures. To avoid tall semantics, we recommend using 3x3 extended convolutions [3] instead of the first module's 5x5 convolution operation. The idea of using wider convolution operations with both the initiation module as a severe way to deal with multi-scale data can be found in [2], but here recently added 1x1 convolutions for measurement reduction.

- Brief (Binary robust independent elementary feature)

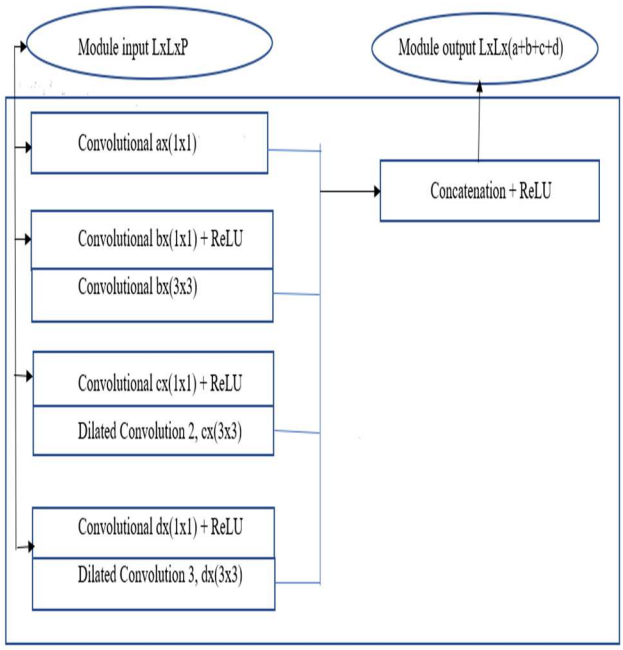

Fig 3: Structure of the inception modules applied in inception net

Here the figure shows how step by step process has happened in the inception net. Starting from module input LxLxP, it is processed layer by layer and at last, it is concatenated and gives module output LxLx(a+b+c+d) respectively.

Here the dataset that is generated is its own data file because, to our knowledge, none exists that compiles recording produced with deepfake technology. The fraudulent assignment requires many days of preparations using conventional processors in order to achieve realistic results, while it is theoretically possible to prepare for two particular faces at once. In order to have a large enough selection, to collect the many clips that are freely available to the public online.

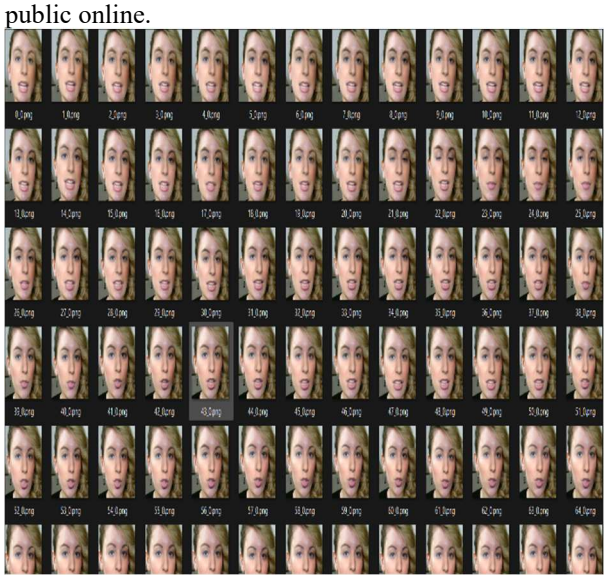

Fig 4: Training data file

A predefined neural arrangement for localizing the facial center of interest in a video based on how many camera angles and light changes the object is exposed to. In

comparison, each scenario necessitated the extraction of approximately 50 faces. Since then, real-face photographs extracted from other websites with the same resolutions have been added to the collection. The confrontation has now been physically examined in order to correct misalignment and off-base placement. To avoid bias in the classification task, an identical set of high- and low-quality figures was used in dual classes as much as possible.

C. AGGREGATION ON INTRA-FRAMES

The paper performed the same image collection that outlines that aren't generated by time, to determine that a lower number of compression artifacts will aid in improving the stratification marks. This allowed us to gain a better understanding of how video compression affects imitation location. On the other hand, films that are only a few seconds long may only include three intra frames, negating aggregate, discovered may have terrible consequences in the categorization, but even so, the distinctiveness is little, because the generated scores are greater than zero and it can be used as a quick conglomeration.

D. FACE TO FACE DATA FILE

We have investigated if the suggested approach may be used to identify other confront frauds in addition to the deepfake data file. This Face Forensics data file [20] is a strong contender since it includes over a thousand fake recordings that were made specifically using the Face2Face method. This data file is currently a component of a collection for preparation, approval, and testing. One benefit of the Face Forensics collection is the availability of losslessly compressed movies, which has allowed us to examine the strength of their model at various compression settings rather than just enhancing the use of the suggested architecture to some other classification job. We picked the same H.264 compression rate with compression, 23 (moderate compression).

E. NETWORK INTUITION

Here efforts were made to know why such systems address the categorization issue. By analyzing the mass of the many C- neurons & parts as figure descriptors, this issue might be accomplished. The collection of positive, negative, and then gain weights can be read as the derivation of a discrete moment arrangement. In case of faces, this is less significant because it only applies to the main layer. Another way is to produce a gain picture that increases the enactment to a particular. Simply put, the yield of channel j of layers I and x is seen and it is the maximizing of looks such that it is the most severe enactment for shown to of the final covering layer of Meso4 which ultimately comes down to the faja irregular image, and it places a regularization on the output to decrease noise. Thus, prepared to segregate those neurons according to the direction of the mass keeping track of their behavior gaining forward towards a refusing rating when compared to fabricated course or a positive score when compared to the real instruction.

Contrastingly, negative-weighted neurons reveal powerful subtle aspects on the foundation section, clearing away a smooth face area, while gained-weighted neurons show images with profoundly eyes, nose, and mouth regions that make sense because deepfake-generated faces sometimes lack delicate details or blurry in comparison to the left of the image, which was rest intact.

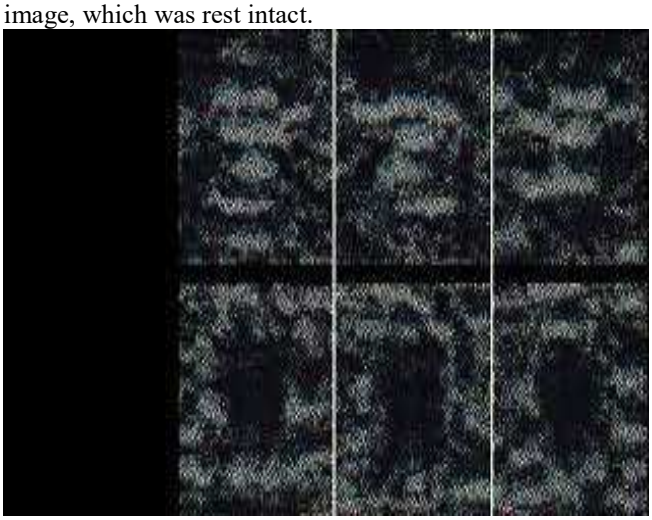

Fig 5: Activation of neurons of layer which is hidden

Furthermore, it can analyze harsh layer results for batches of real and fake figures, and it can show the difference in actuation and identify the elements of the gain in the figure that is important for categorization. Let us consider the trained inception setup on the deepfake data file, which is significantly stimulated for real photographs but not for deepfake pictures where the backdrop displays the most notable crests.

Thus, it concludes that it is once again a matter of fuzziness, with the eye being perhaps the minute detail in original photographs & foundation in manufactured photographs due to the size decrement suffered by the human face.

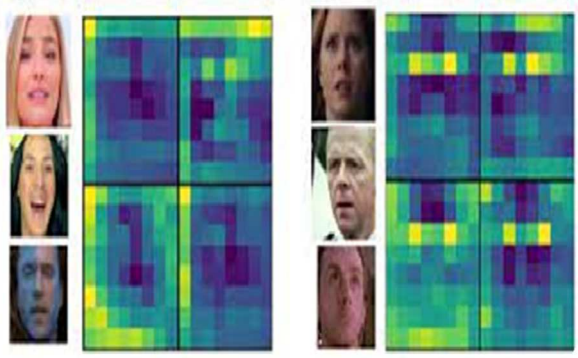

Fig 6: Mean layer output of 100 deepfake faces and Mean layer output of 100 real faces

Here figure 6 shows the output from the convolutional neural network (CNN data file for several of the filters in inception-4's (fourth version of the inception model of deep CNN final neural layer)

Here this figure also shows the mean layer output of 100 deepfake faces and the mean layer output of 100 real faces respectively. It shows faces that are manipulated and which are not manipulated. Here from this figure, the faces which are real and which are fake are done by inception net respectively.

F. INCEPTION NET ARCHITECTURE

A starting arrangement might be a complex neural arrangement with a structure made up of repetition units referred to as initiation modules. As it was previously said, the focus of this article is the initiation module's specific subtle features.

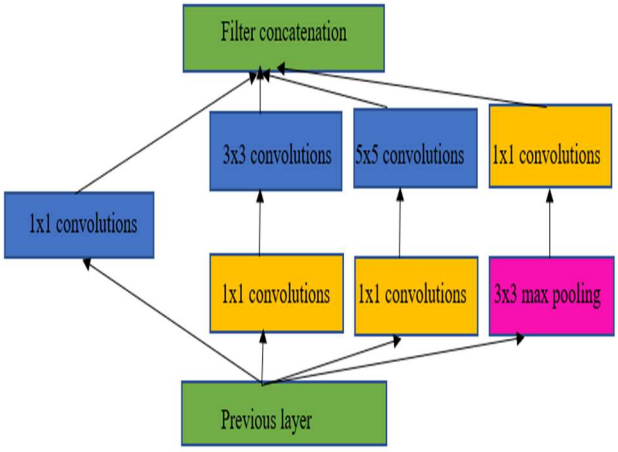

Fig 7: Inception net architecture

Here figure 7 shows how this architecture works and this is the block representation of the inception net. Here different layers are used such as convolution and max pooling that is 1x1,5x5 and 3x3 convolution layers and 3x3 max pooling layer.

IV. RESULTS AND ANALYSIS

The dataset which has been used in this paper is the face forensics dataset and the Deepfake detection challenge(DFDC) dataset. This dataset contains 470 GB of videos which are 124000 videos in total for DFDC and 30GB of videos which are 5000 videos in total for the face forensics dataset. Different versions of the dataset are produced for every step.

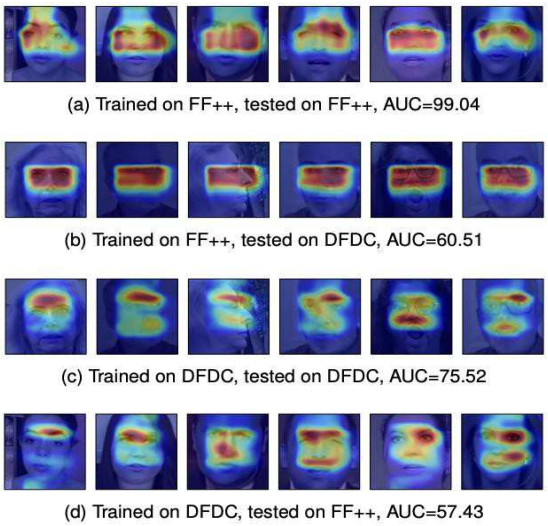

Fig 8: Dataset: Version 0 initial dataset to Version 3 final dataset

Tests using face forensics++ were also conducted in terms of comparing various produced images of different data sets. With different sub-datasets of face forensic++, except Deep fakes, the used architecture is better than standard conventional architectures. This is likely a result of the network's improved ability to generalize about extremely particular deepfakes.

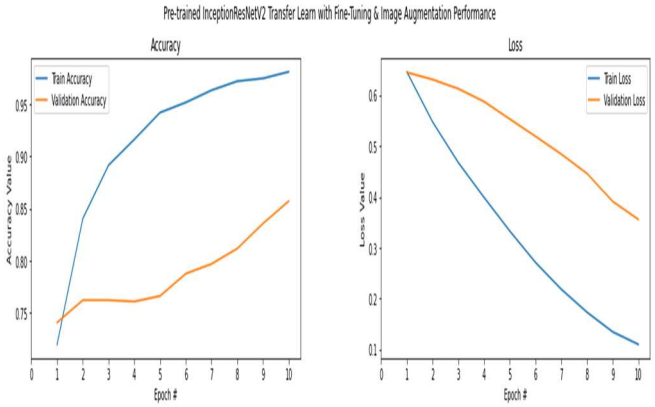

Fig 9: Accuracy and Image augmentation performance (Loss)

Figure 9 shows the accuracy and image augmentation performance(loss) of the pre-trained inception net with fine-tuning. here it is possible to see the accuracy of a dataset by using inception net.

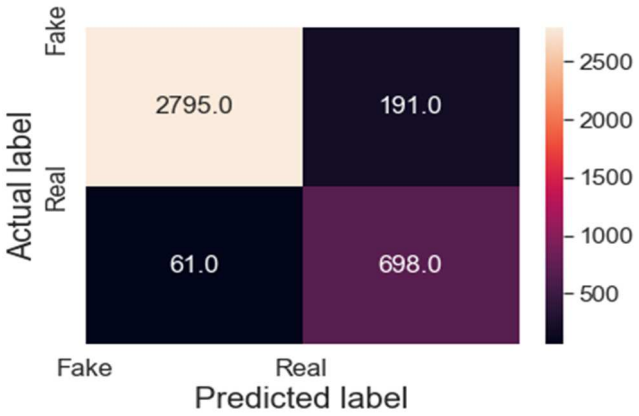

Fig 10: Confusion matrix

Here figure 10 shows how the confusion matrix is got for the DFDC dataset. It also shows how actual label and predicted label is done with real and fake images. It shows real and fake images with digits as shown in figure 10.

Since a confusion matrix describes that it is a matrix of numbers which shows that a model gets confused. It is a class-wise distribution of a classification model's predictive performance—that is, the confusion matrix is an organized way of mapping the predictions to the original classes to which the data belongs.

The confusion matrix shown in figure 10 is an example of one that is obtained with a trained model for the DFDC dataset. It provides more information than just the model's accuracy.

Here the figure gives a total of 2795+61=2856 positive samples by adding the numbers in the first column. Similarly, it gives the number of samples in the negative class by adding the numbers in the second column, which is 889. The total

number of samples evaluated is represented by the addition of numbers in the overall boxes. Furthermore, the diagonal elements of the matrix are correct classifications that are 2795 for the positive class and 698 for the negative class.

The model now classified 61 samples (bottom-left box) that were expected to be positive as negative. So, it is called "False Negatives" because the model predicted "negative," which was incorrect. Similarly, 191 samples (top-right box) were expected to be negative but were classified as "positive" by the model. As a result, they are known as "False Positives." By using the four different numbers from the matrix, it is easier to evaluate the model more thoroughly.

V. CONCLUSION

In this work, the InceptionNet architecture has been used for identifying the fake faces. Different types of transitions in real images with test parameters, such as the number of key points in images, comparison rate, and performance time required for each algorithm are used. This paper shows overall accuracy for the DFDC dataset as 93%. This work can classify deepfakes recordings from various resources with diverse convolutional layers. Thus, this paper's contribution will inevitably help with the diminishment of fake recordings and coercion in our society. The proposed work was completed more faster than the existing work, and the detection of fake and real images was very effective. In the DFDC dataset, the accuracy rate of proposed work reached 93%. It could be extended in the future to use different classifiers and distance metric measures to detect deepfake face images.

VI. REFERENCES

[1] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.

[2] Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. "Inceptionism: Going deeper into neural networks." (2015).

[3] Badale, Anuj, et al. "Deep fake detection using neural networks." 15th IEEE international conference on advanced video and signal-based surveillance (AVSS). 2018.

[4] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[5] Bayar, Belhassen, and Matthew C. Stamm. "A deep learning approach to universal image manipulation detection using a new convolutional layer." Proceedings of the 4th ACM workshop on information hiding and multimedia security. 2016.

[6] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

[7] Chen, Chun-Fu Richard, Quanfu Fan, and Rameswar Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[8] Heo, Young-Jin, et al. "Deepfake detection scheme based on vision transformer and distillation." arXiv preprint arXiv:2104.01353 (2021).

[9] Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE signal processing letters 23.10 (2016): 1499-1503.,

[10] Kaggle,https://www.kaggle.com/competitions/deepfake-detection-challenge/data